BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 9: e183181
doi: 10.3897/biss.9.183181

OPEN ACCESS

Conference Abstract

Retrieving Information about Taxonomic Expertise from Specimen Data and the Importance of Data Quality for Effective Data Reuse

David Fichtmueller‡, Sofie Meeus§, Quentin Groom§, Anton Güntsch‡

‡ Freie Universität Berlin, Botanic Garden and Botanical Museum, Berlin, Germany
§ Meise Botanic Garden, Meise, Belgium

Corresponding author: David Fichtmueller (d.fichtmueller@bgbm.org)

Received: 21 Dec 2025 | Published: 23 Dec 2025

Citation: Fichtmueller D, Meeus S, Groom Q, Güntsch A (2025) Retrieving Information about Taxonomic Expertise from Specimen Data and the Importance of Data Quality for Effective Data Reuse. Biodiversity Information Science and Standards 9: e183181. https://doi.org/10.3897/biss.9.183181

Abstract

Understanding the current landscape of taxonomic expertise is crucial for addressing the growing challenges in biodiversity research, conservation, and data management. There remains a lack of comprehensive insight into who the active taxonomists are, where they work and which taxon groups they specialize in. Within the European TETTRIs project (Transforming European Taxonomy through Training, Research and Innovations) we aim to fill that gap by systematically determining the taxonomic expertise in Europe, using different sources. De Nolf et al. (2023) showed how to retrieve expertise from taxonomic publications.

This talk focuses on an approach to retrieve taxonomic expertise from specimen data. It will highlight how data quality issues can be a significant hurdle for effective data reuse. We consider a current European expert as anyone who has either collected or identified a specimen in the last 10 years and who is working in or coming from a country of the European Political Community (EPC).

© Fichtmueller D et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The core workflow of the specimen approach is shown in Fig. 1, however for this talk we want to focus on the first three steps.

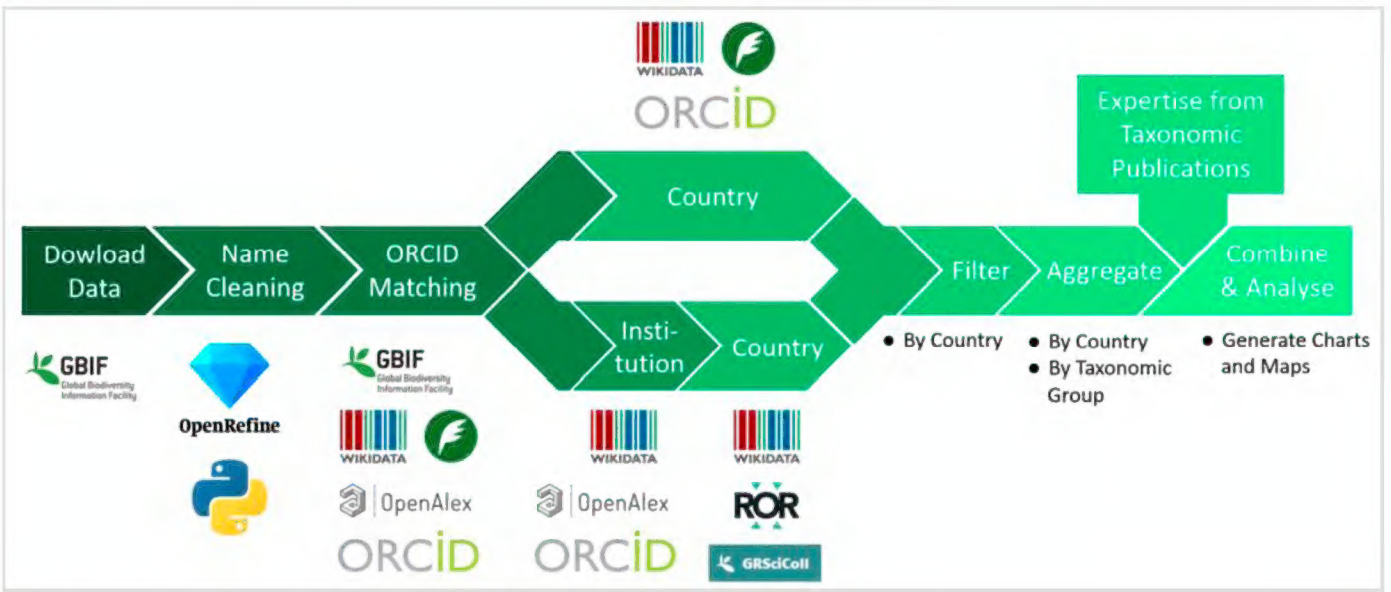

Figure 1.

Processing workflow for determining taxonomic expertise from specimens. (CC-BY David Fichtmueller)

The core data consists of around 20.5 million specimen records retrieved from GBIF via the new SQL-API in the DarwinCore (DwC) format (GBIF.org 2025). From this dataset, the columns for names and identifiers of the people who collected or identified the specimen (dwc:recordedBy and dwc:recordedByID, as well as dwc:identifiedBy and dwc:identifiedByID respectively), were extracted, combined and aggregated, as is demonstrated in Fig. 2. This reduced the number of records to 570,000 unique name and identifier combinations.

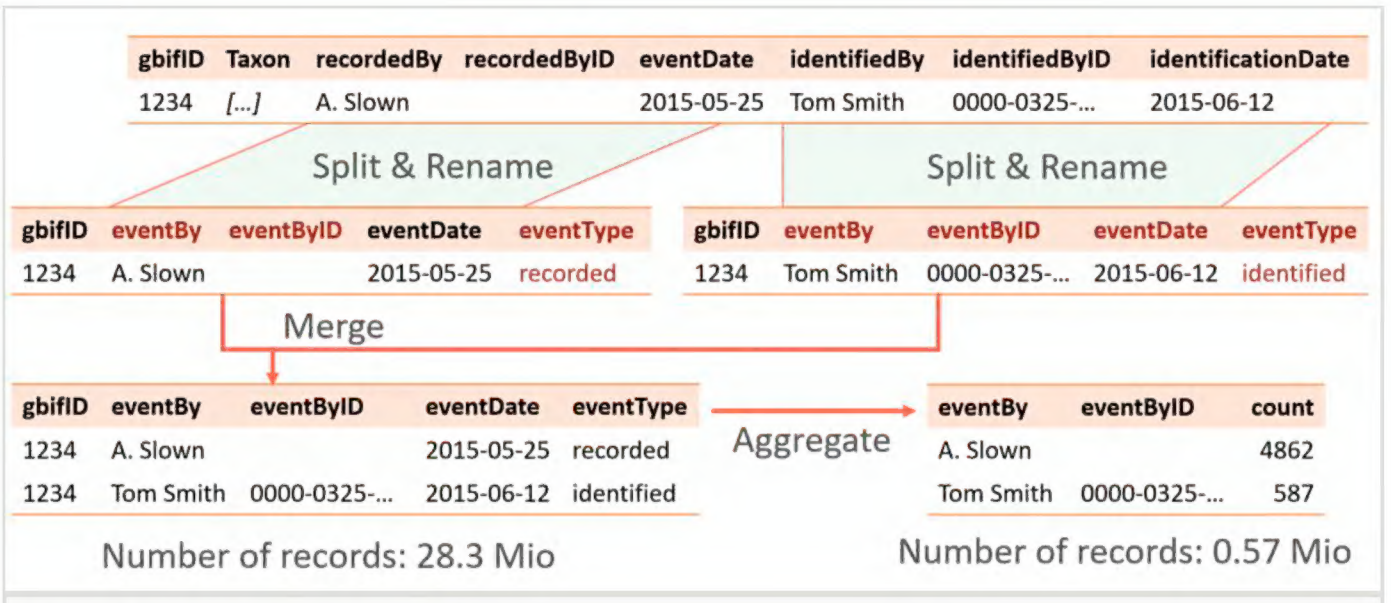
gbifID Taxon recordedBy recordedByID eventDate identifiedBy identifiedByID identificationDate
1234 [...] A. Slown 2015-05-25 Tom Smith 0000-0325-... 2015-06-12
Split & Rename
Split & Rename
gbifID eventBy eventByID eventDate eventType
1234 A. Slown 2015-05-25 recorded
gbifID eventBy eventByID eventDate eventType
1234 Tom Smith 0000-0325-... 2015-06-12 identified
Merge
gbifID eventBy eventByID eventDate eventType
1234 A. Slown 2015-05-25 recorded
1234 Tom Smith 0000-0325-... 2015-06-12 identified
Aggregate
eventBy eventByID count
A. Slown 4862
Tom Smith 0000-0325-... 587
Number of records: 28.3 Mio
Number of records: 0.57 Mio

Figure 2.

Combining names and person identifiers for collection and identification events from GBIF.org. (CC-BY David Fichtmueller)

Only around 2% of the name/identifier combinations actually had an identifier associated with them. This limited the usability of the data for the next processing steps as any

subsequent findings would be based on only this small subsample. Thus, further cleaning, disambiguation and matching of names became necessary.

The biggest obstacle during the cleanup phase were huge inconsistencies with which names where written. In DwC, the fields dwc:recordedBy and dwc:identifiedBy are not repeatable, therefore it is common to have the names of multiple persons in each field. The recomended way of concatenating multiple records using the vertical bar symbol (" | ") is rarely used. More often, enumerations are made using commas, semicolons, the ampersand symbol, or the word "and" in various languages. Other issues were the inconsistent ordering and separation of given and family names as well as group attributions or non-name content, such as numbers, affiliations or roles.

Further processing was done using a Python script that took the different formats, rules and exceptions into account, to split the combined names into individual names and then their given and family name parts, reducing the number of records to 430,000. At this point we would like to acknowledge that many of the processing steps rely on generalization and heuristics that will not do justice to many names, as well as our strong emphasis on European name customs. We are aware of these shortcomings and recognize that people's names are more complex than this approach is able to handle, as shown by the insightful and entertaining article "Falsehoods Programmers Believe About Names" (McKenzie 2010). Despite this, we still consider our general conclusions as valid, as the process relies on knowledge aggregation.

The cleaned-up names could then be better matched to other GBIF records, Bionomia, Wikidata and matches from the taxonomic publications approach. While the results of the subsequent analysis are beyond the scope of this talk, the necessary preprocessing steps clearly showed that data coming from different sources is messy and requires extensive cleanup in order to properly reuse it.

We encourage people to leverage their roles in the biodiversity research community to improve data and foster easy data reuse. Individual researchers, software developers and journal editors and reviewers can all play a part, e.g., by using and promoting ORCIDs, improving data entry mechanisms, or warning about data quality issues and inconsistencies. Together, we can improve data quality and data consistency in our field, as they are key factors for effective data reuse across research projects.

Keywords

person names, workflow, GBIF, recordedBy, identifiedBy

Presenting author

David Fichtmueller

Presented at

Living Data 2025

Grant title

TETTRIs, Grant Agreement 101081903

Conflicts of interest

The authors have declared that no competing interests exist.

References

- De Nolf M, Meeus S, Fichtmueller D, Lundin K, Tilley L, Groom Q (2023) European Taxonomists in Profile: A Data-Driven Approach. Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.111534
- GBIF.org (2025) GBIF Occurrence Download. Release date: 2025-5-14. URL: https://doi.org/10.15468/dl.2jyxx8
- McKenzie P (2010) Falsehoods Programmers Believe About Names. https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/. Accessed on: 2025-10-31.